ISCA Archive
http://www.isca-speech.org/archive

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

A Novel Feature Extraction Using Multiple Acoustic Feature Planes for HMM-based Speech Recognition

Tsuneo NITTA, Masashi TAKIGAWA and Takashi FUKUDA
Graduate School of Eng., Toyohashi University of Technology
1-1 Hibariga-oka, Tempaku, Toyohashi JAPAN E-mail: nitta@tutkie.tut.ac.jp

ABSTRACT

This paper describes an attempt to extract multiple peripheral features of a point $\mathbf{x}(t_i,f_j)$ on a time-spectrum (**TS**) pattern by observing n×n neighborhoods of the point, and to incorporate these peripheral features (**MPFPs**: multiple peripheral feature planes) into the feature extractor of a speech recognition system together with **MFCC** parameters. Two types of peripheral feature extractor, **MPFP-KL** and **MPFP-LR**, are proposed. **MPFP-KL** adopts the orthogonal bases extracted directly from speech data by using **KLT** of 7×7 blocks on **TS** patterns. In **MPFP-LR**, the upper two primal bases are selected and simplified in the form of Δ_t-operator and Δ_f-operator obtained by linear regression calculation. **MPFP-KL** and **MPFP-LR** show significant improvements in comparison with the standard **MFCC** feature extractor in experiments with the **HMM**-based ASR system.

1. INTRODUCTION

Time-spectrum (**TS**) pattern $\mathbf{x}(t,f)$ has long been used for acoustic features in automatic speech recognition (ASR), and recently, dynamic features such as Δ-cepstrum, Δ-power, etc. have been introduced into ASR [1],[2] and the set of **MFCC** and dynamic features is widely used. Dynamic features represent peripheral features of a point on a **TS** pattern $\mathbf{x}(t_i,f_j)$ along the time axis, however, we can obtain more information from n×n neighborhoods of the point. In this paper, we investigate primal peripheral features embedded in n×n blocks of **TS** patterns first.

In the previous work [3], the feature extraction method based on multiple acoustic feature planes (**MAFPs**) was applied to a phonetic segment classification task and showed that the method significantly improved the error rate. In the method, a set **X** with elements $\mathbf{x}(t, f)$ is mapped onto multiple **AFPs** (acoustic-feature planes) $\mathbf{Y}_m=\mathbf{y}_m(t, f)$, m=1, 2, .,.M by using local mapping operators $\{\mathbf{G}_m\}$ ($\mathbf{G}_m \in \mathbf{G}$):

$$\mathbf{G}_m : \mathbf{X} \rightarrow \mathbf{Y}_m \quad (1)$$

Firstly, to obtain the $\mathbf{G}_m$, the 3×3 orthogonal basis $\{\mathbf{F}_1, \mathbf{F}_2, ..., \mathbf{F}_9\}$ on **TS** patterns was extracted directly from a speech database. Then, the orthogonal basis was replaced to a set of modeled operators that simplified $\{\mathbf{F}_m\}$ and made them symmetrical. In this paper, we try not to extract multiple local feature planes (**MLFPs**) but to extract multiple peripheral feature planes (**MPFPs**), then incorporate **MPFPs** into an **HMM**-based ASR system together with **MFCC** parameters through the same approach used in the **MLFP** extraction process [3].

This paper is organized as follows: Section 2 discusses the geometrical structure of 7×7 blocks on **TS** patterns. Section 3 then outlines methods of implementing the peripheral features in a feature extractor of an ASR system together with **MFCC** parameters. Finally, Section 4 gives the experimental setup, the results and discussion.

2. OBSERVING PERIPHERAL FEATURES ON TS PATTERNS

We can observe many types of geometrical structures on **TS** patterns. **Figure 1** shows the upper nine elements of an orthogonal basis of 7×7 blocks on **TS** patterns analyzed by a BPF bank with 24 channels. The 7×7 orthogonal basis was extracted by using Karhunen-Loeve transform (**KLT**) from speech data described in section 4.1.

From a space-operational point of view, $\mathbf{F}_1$ is considered to be a smoothing operator and this neutral operator generally has no effect on feature extraction for ASR. $\mathbf{F}_2$ and $\mathbf{F}_3$ are the first-order derivative operators with respect to the time axis (Δ_t-operator) and frequency axis (Δ_f-operator), respectively, $\mathbf{F}_4$, $\mathbf{F}_9$ are the second-order derivative operators with respect to the time axis ($\Delta_t\Delta_t$-operator) and frequency axis ($\Delta_f\Delta_f$-operator), respectively, and $\mathbf{F}_5$, $\mathbf{F}_6$, $\mathbf{F}_7$, $\mathbf{F}_8$ are subspaces that represent ridges and/or valleys on **TS** patterns.

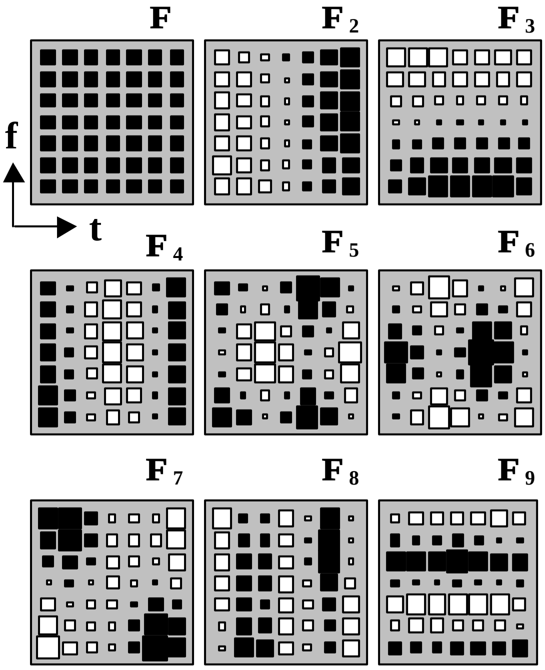

Figure 1 7´ 7 orthogonal basis extracted from speech data

Time-frequency space operators {$\mathbf{F}_m$}, or mapping operators {$\mathbf{G}_m$}, map a **TS** pattern $\mathbf{x}(t, f)$ onto multiple **PFP**s (peripheral feature planes) $\mathbf{Y}_m = \mathbf{y}_m(t, f)$, m=1, 2, .,.M. An element $y_m(t, f)$ of **MPFP** is calculated with 7×7 neighborhoods of x(t, f) and $\mathbf{G}_m = \mathbf{g}_m(t, f)$ by the following equation:

$$\mathbf{y}_m(t, f) = \sum_{i=-3}^{3} \sum_{j=-3}^{3} \mathbf{x}(t+i, f+j)\, \mathbf{g}_m(i, j) \qquad (2)$$

Figure 2 shows an example of the upper three **PFP**s of an utterance [kaden' tsa] (cadence). In the figure, (A) is an original **TS** pattern and (B), (C), and (D) represent the 2nd-**PFP** mapped with a Δ_t-operator $\mathbf{F}_2$ ($\mathbf{G}_2$), the 3rd-**PFP** mapped with a Δ_f-operator $\mathbf{F}_3$ ($\mathbf{G}_3$), and the 4th-**PFP** mapped with $\Delta_t\Delta_t$-operator $\mathbf{F}_4$ ($\mathbf{G}_4$), respectively. A positive sign of $\mathbf{y}_m(t, f)$ means a positive slope, a negative sign a negative slope. For example, a clear spectral peak in steady sound is represented by a pair of positive and negative values in the 3rd-**PFP**. In the figure, patterns on **PFP** are displayed with absolute values.

3. COMBINING PERIFERAL FEATURES WITH MFCC-PARAMETERS

This chapter describes the methods of extracting peripheral features and combining them with **MFCC**-parameters in a feature extractor. **Figure 3-A** shows a standard feature parameters used in current **HMM**-

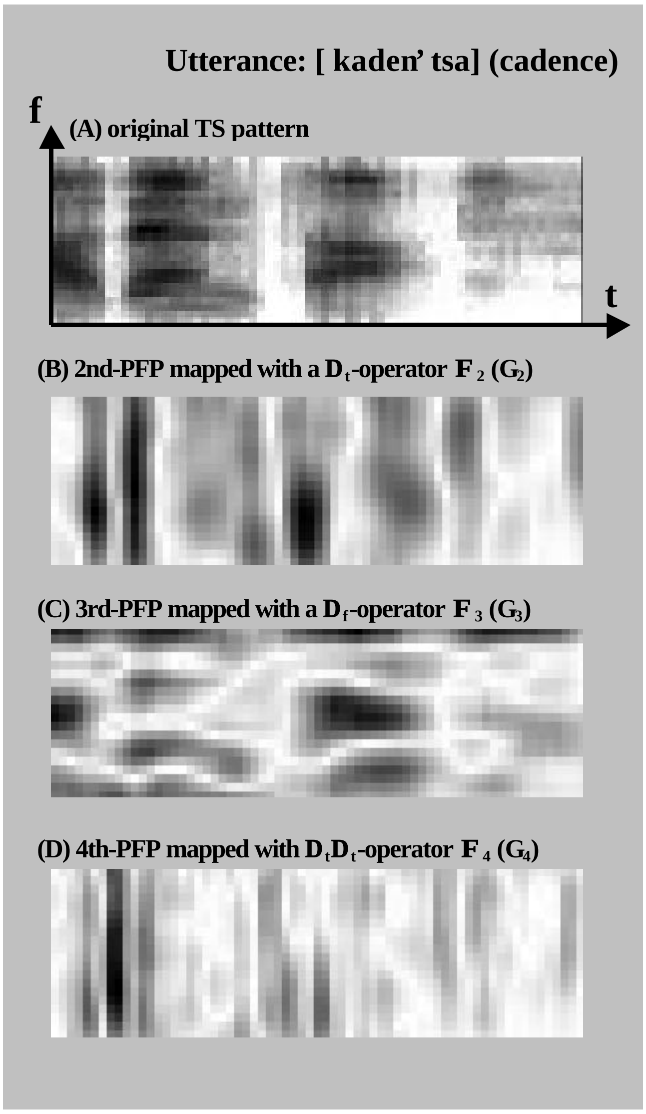

Figure 2 Time-spectrum pattern and peripheral feature planes (PFPs).

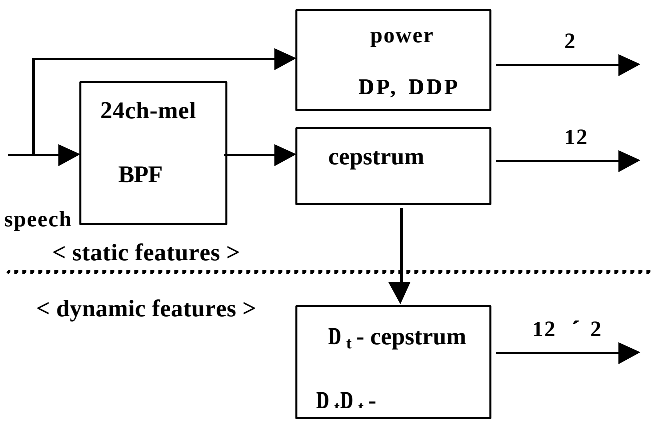

Figure 3-A MFCC with dynamic features : baseline

based ASR systems. In the feature extractor, an input speech is sampled at 16 kHz and a 512-point **FFT** of the 25 ms Hamming-windowed speech segments is applied every 10 ms. The resultant **FFT** power spectrum is then integrated into the output of 24ch-**BPFs** that have mel-scaled center-frequencies. Then, 38 feature parameters including 12 static parameters (mel-cepstrum), ΔP (logarithmic power), ΔΔP, and 24 dynamic features (Δ_t, $\Delta_t\Delta_t$) are extracted after converting the output of **BPFs** into cepstrum coefficients (**MFCCs**).

Figure 3-B shows the procedure of extracting peripheral features. In the figure, firstly, an output of 24ch-**BPF** bank $\mathbf{x}(t,f)$ is mapped onto **PFPs** $\mathbf{y}_m(t,f)$, m=1,2,..,8; f=1,2,..,20 by equation (2). Next, each **PFP** is converted into cepstrum coefficients $\mathbf{c}(m,q)$, m=1,2,..8; q=1,2,..10 by **DCT**. Finally, $\mathbf{c}(m,q)$ with 80 dimensions are compressed into a selected peripheral-feature-vector $\mathbf{z}$ (k), k=1,2,..,24 through **KLT** by the following equation:

$$\mathbf{z}(k) = \sum_{m=1}^{8} \sum_{q=1}^{10} \mathbf{c}(m,q)\ \boldsymbol{\varphi}_k(m,q) \qquad k=1,\ 2,..,\ 24 \qquad (3)$$

where, $\boldsymbol{\varphi}_k(m,q)$ is the k-th eigen vector set of **KLT**. 24 peripheral features $\mathbf{z}(k)$ are combined with **MFCC** static features (12 **MFCC** + ΔP, ΔΔP).

In chapter 2, we investigated the 7×7 orthogonal basis on **TS** patterns and found that the upper two primal bases were Δ_t-operator and Δ_f-operator. On the other hand, the standard feature vector set of **MFCC**-based parameters did not include Δ_f related ones. **Figure 3-C** shows the other type of peripheral feature representation. In this figure, two space operators that give two peripheral features of Δ_t- and Δ_f- cepstrum are simplified in the form of 7×1-block operator (Δ_t) and 1×7-block operator (Δ_f), and the derivative operation is replaced by the calculation of linear regression. 24 peripheral features including 12 Δ_t-cepstrum coefficients, or dynamic features, are combined with **MFCC** static features (12 **MFCC** + ΔP, ΔΔP).

4. EXPERIMENTS

4.1 Speech Database

The following four data sets were used.

D1. Acoustic model design set: A subset of "ASJ (Acoustic Society of Japan) Continuous Speech Database", consisting of 4,503 sentences uttered by 30 male speakers (16kHz, 16 bit).

D2. Test data set: A subset of "Tohoku University and Matsushita Spoken Word Database", consisting of 100 words uttered by 10 unknown male speakers. The

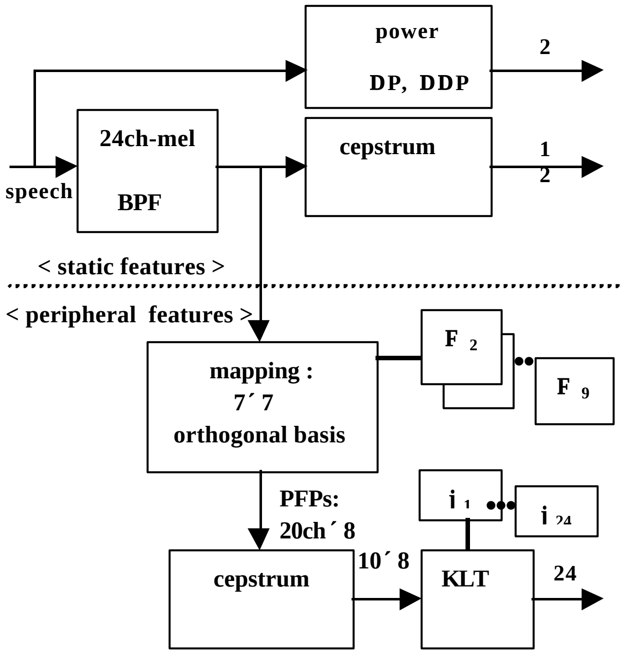

Figure 3-B MFCC with peripheral features : MPFP-KL (7´ 7 orthogonal basis)

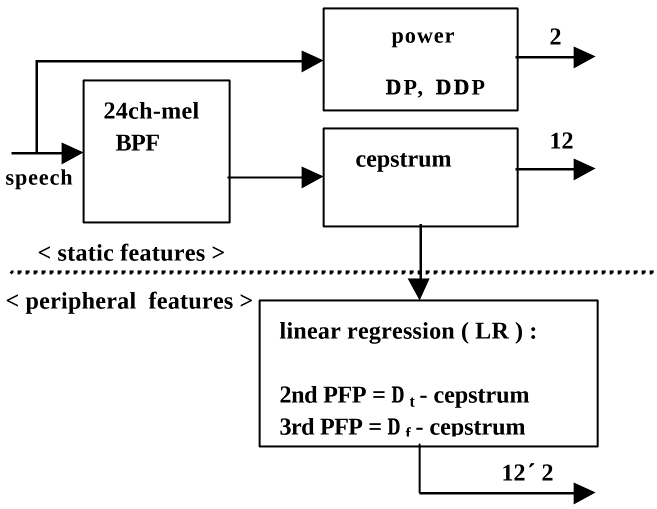

Figure 3-C MFCC with peripheral features : MPFP-LR (D t - & D f – cepstrum)

sampling rate was converted from 24 kHz to 16 kHz.

D3. 7×7 orthogonal basis design data set: A subset of "ASJ News Corpus (ASJ-JNUS)", consisting of 2,662 sentences uttered by 53 male speakers.

D4. Eigen-vectors design set for **KLT**: A subset of the same ASJ-JNUS corpus used for designing the orthogonal basis, consisting of 5,569 different sentences uttered by the same 53 male speakers.

4.2 Experimental Setup

Table 1 shows specifications of the three methods applied in the feature extractor of the experimental ASR system. All the methods use **MFCC** (12), ΔP (1), and ΔΔP (1) and have the same 38 dimensions. Firstly, 43 Japanese monophone-HMMs with five states and three loops are designed with a **D1** data set. In the HMM, output probabilities are represented in the form of a Gaussian mixture and covariance matrices are diagonalized. Next, speaker-independent word-recognition tests are carried out with a **D2** data set.

4.3 Results and Discussion

Table 2 compares the word recognition rates between the three feature extractors. The results show that the recognition scores of **MPFP-KL** with peripheral features are higher than those of the baseline extractor that has only dynamic features. Especially, the score for the one-mixture model of **MPFP-KL** is significantly higher than that of the baseline. This fact suggests that, from a robust feature-extracting point of view, **MFCC** with peripheral features is superior to **MFCC** with dynamic features.

Why does **MPFP-KL** show high performance? In the three orthogonal bases shown in **Figure 1**, $\mathbf{F}_2$ and $\mathbf{F}_4$ were already adopted in the baseline extractor as dynamic features, however, $\mathbf{F}_3$ that shows the upper contribution in **KLT** was not introduced yet. **MPFP-LR** in **Table 2** discards $\Delta_t\Delta_t$-cepstrum from the baseline extractor and adds Δ_f-cepstrum to it substitutingly. The result in **Table 2** shows the importance of adding dynamics along the frequency axis to the standard **MFCC** parameter set. **Figure 4** shows the experimental results when various types of features are added to MFCC.

5. CONCLUSION

A framework for incorporating multiple geometric structures into the feature extractor of ASR systems was proposed. The design methodology of mapping operators for extracting peripheral features was given by observing the orthogonal basis of speech and by incorporating primal components into a feature extractor in a simplified form. The proposed method based on MPFP-KL or MPFP-LR showed significant improvements in comparison with the standard MFCC feature extractor in the experiments with the HMM-based ASR system. It is very important to add dynamics along the frequency axis to the standard **MFCC** parameter set.

REFERENCES

[1] K. Elenius and M. Blomberg, "Effect of emphasizing transitional or stationary parts of the speech signal in a

Table 1 Three methods for feature extraction

method	parameters	dimension
Baseline	**MFCC + D_t + D_tD_t + DP+ DDP**	**38**
MPFP-KL	**MFCC + KL + DP+ DDP**	**38**
MPFP- LR	**MFCC + D_t + D_f + DP+ DDP**	**38**

Table 2 Comparison for three feature extractors

method	word correct rate [%]		
	mix. = 1	mix. = 2	mix. = 4
Baseline	**93.58**	**93.99**	**95.82**
MPFP-KL	**97.25**	**97.15**	**97.15**
MPFP- LR	**97.96**	**98.37**	**98.47**

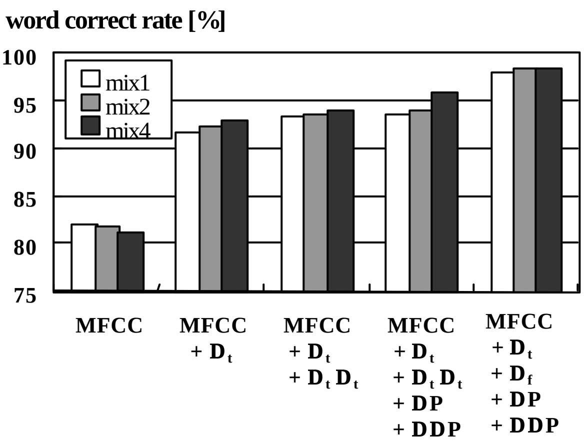

Figure 4 Comparison between MFCC parameter sets.

discrete utterance recognition system", IEEE Proc. ICASSP' 82, pp.535-538 (1982).
[2] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum", IEEE Trans. Acoust. Speech Signal Process. ASSP-34, pp.522-59 (1986).
[3] T. Nitta, "Feature extraction for speech recognition based on orthogonal acoustic-feature planes and LDA," Proc. IEEE ICASSP' 99, Phoenix, Vol.1, pp.421-424 (1999-3).